Regression and Balancing

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We introduced another method to account for the influence of confounders, weighting.
- It is based on the argument of Rosenbaum and Rubin that the propensity score is central for causal inference.
- It contains all the information from the confounders and should be a balance score.
- We can estimate the propensity scores with either logistic regression or the CBPS method.
- Then, we rely on the IPW estimators (HT or HA) to obtain the estimates.
- Using the estimated propensity scores is more efficient than using the true propensity scores.
- Ignoring the uncertainty from estimating the propensity scores leads to conservative variance estimates.

Regression

- Conventionally, we use regression models to control for the influence of confounders.
- We run the following regression:

$$Y_i = \tau D_i + \mathbf{X}_i' \beta + \varepsilon_i.$$

- Except for strong ignorability, we also assume that

$$E[Y_i | D_i, \mathbf{X}_i] = \tau D_i + \mathbf{X}_i' \beta$$

- The impacts from the regressors are additive, linear, and homogeneous.
- These are structural restrictions that are usually unjustified.

Regression: caveats

- Remember that when D_i is randomly assigned, controlling $\mathbf{X}_i$ does not cause any bias asymptotically.
- This is no longer the case when we implement block randomization or assume strong ignorability.
- Linearity is more acceptable since we can control for high-order terms in the regression.
- Let's assume that

$$
\begin{aligned}
Y_i(0) &= \mathbf{X}_i'\beta, \\
Y_i(1) &= Y_i(0) + \tau_i, \\
E[D_i|\mathbf{X}_i] &= \mathbf{X}_i'\eta.
\end{aligned}
$$

- Aronow and Samii (2016) show that under these conditions,

$$\hat{\tau}_{OLS} \rightarrow \frac{E[w_i \tau_i]}{E[w_i]} \neq E[\tau_i],$$

where $w_i = (D_i - E[D_i|\mathbf{X}_i])^2$.

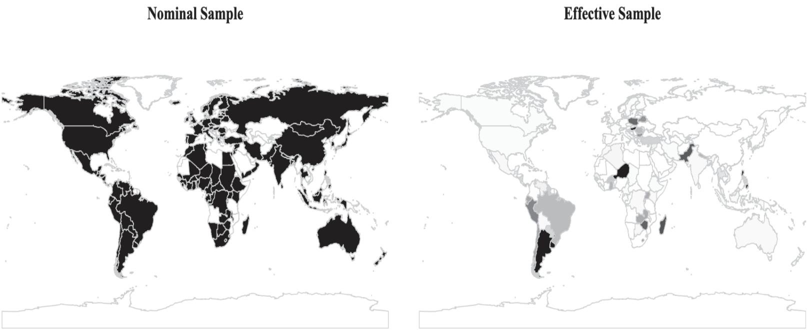
Regression: caveats

- The OLS estimate converges to a "convex combination" of individualistic treatment effects.
- In the definition of the SATE, each τ_i has a weight of $\frac{1}{N}$.
- But in the OLS estimate, the weights vary across the units.
- In general, $\hat{\tau}_{OLS}$ does not converge to τ unless $\tau_i = \tau$ (homogeneity).
- Observations have unequal contributions to the estimate.
- The group in which the treatment varies more drastically is over-weighted in the analysis.
- Thus, the OLS estimate is not representative of the sample and does not have a higher external validity.
- Aronow and Samii (2016) define the concept of "effective sample," the sample re-weighted by each unit's contribution to the estimate.

Regression: caveats

In the effective sample, each unit has a weight of $\hat{\nu}_i^2$, where $\hat{\nu}_i$ is the residual from regressing D_i on $\mathbf{X}_i$.

FIGURE 1 Example of nominal and effective samples from Jensen (2003)



Note: On the left, the shading shows countries in the nominal sample for Jensen (2003) estimate of the effects of regime type on FDI. On the right, darker shading indicates that a country contributes more to the effective sample, based on the panel specification used in estimation.

Regression: caveats

- Regression may prevent you from seeing the failure of positivity.
- Let's assume that $\tau_i = \tau$, and fit regression models on both the treatment group and the control group.
- The regression coefficients will be consistently estimated, and we can calculate the predicted outcome $\hat{Y}_i$ for any unit.
- Then, we estimate τ by

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \hat{Y}_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - D_i) \hat{Y}_i.$$

- It is straightforward to show that

$$E[\hat{\tau}] \to \tau + (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)\beta.$$

- If positivity fails, $\bar{\mathbf{X}}_1 \neq \bar{\mathbf{X}}_0$ and the estimator is inconsistent.

Regression: remedy

- One solution is to rely on the "counterfactual estimator" (Kline 2011; Heckman, Ichimura, and Todd 1997).
- We first estimate β using only units in the control group.
- Then, we predict the counterfactual of each treated unit via

$$\hat{Y}_i(0) = \mathbf{X}_i' \hat{\beta}.$$

- We estimate τ via

$$\hat{\tau}_{reg} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \left(Y_i - \hat{Y}_i(0) \right) \to \tau_{ATT} = \frac{1}{N_1} \sum_{i:D_i=1} \tau_i.$$

- This estimator does not suffer from the asymptotic bias as we weight the treated units properly.
- The idea is later generalized to "X-learner" by Künzel et al. (2019) where they use machine learning to predict $\hat{Y}_i(0)$.

Regression: pros and cons

- Regression requires the correct model specification and relies on extrapolation when positivity fails.
- We can increase the complexity of the outcome model to reduce the bias caused by treatment effect heterogeneity (Ratkovic 2019).
- For instance, with one confounder, we can fit two kernel regression models, $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$, on the treatment group and the control group, respectively.
- Then,

$$\hat{\tau}_{reg} = \frac{1}{N}\sum_{i=1}^{N}[\hat{m}_1(X_i) - \hat{m}_0(X_i)]$$

  will be consistent for τ.
- Regression requires weaker assumptions: $E[\varepsilon_i|\mathbf{X}_i] = 0$.

Regression: application

```
## The OLS estimate is 1794.343

## The SE of OLS estimate is 670.9967

## The Lin regression estimate is 1583.468

## The SE of Lin regression estimate is 678.0574
```

Regression: application

```
## The regression ATT estimate is 687.8221
```

Balancing

- Now, consider the outcome model

$$Y_i(0) = \mathbf{X}_i' \beta_0,$$
$$Y_i(1) = \mathbf{X}_i' \beta_1 + \tau_i.$$

- Suppose we can find a group of weights, $\{w_i\}_{(i:D_i=0)}$, such that

$$\bar{\mathbf{X}}_1 = \sum_{i:D_i=0} w_i \mathbf{X}_i, \text{ and } \sum_{i:D_i=0} w_i = 1$$

- Then,

$$\bar{Y}_1 - \bar{Y}_0^w = \tau_{ATT} + (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_1)\beta = \tau_{ATT},$$

  with $\bar{Y}_0^w = \sum_{i:D_i=0} w_i Y_i$.
- This is an idea known as balancing.

Entropy balancing

- There are many possible sets of $\{w_i\}_{(i:D_i=0)}$ that satisfy the balance conditions.
- We choose the set that minimizes a pre-specified criterion, such as the entropy (Hainmueller 2012):

$$\sum_{i:D_i=0} w_i \log w_i.$$

- This method is thus known as "entropy balancing."
- Entropy measures the "uncertainty inherent to the variable's possible outcomes" (Wikipedia).
- It has a root in statistical thermodynamics and was introduced by Shannon when he founded the information theory.
- The weights can be solved via convex optimization.

Entropy balancing

- We estimate the ATT via

$$\hat{\tau}_{EB} = \frac{1}{N_1} \sum_{i:D_i=1} Y_i - \sum_{i:D_i=0} \hat{w}_i Y_i$$

- $\hat{\tau}_{EB}$ is consistent when either the outcome is linear in $\mathbf{X}_i$ or the propensity score model is logistic (Zhao and Percival 2016).
- We can try to balance higher order moments of $\mathbf{X}$.
- The more moments we balance, the more likely we eliminate all the influences of $\mathbf{X}$.
- There can be too many choices if $\mathbf{X}$ is also high-dimensional.
- We need approaches to select moments that matter.
- This can be done by either kernel balancing (Hazlett 2018) or hierarchically regularized entropy balancing (Xu and Yang 2021).

Kernel balancing

- In kernel balancing, we calculate the kernel distance between each pair of units.
- Suppose we use the Gaussian kernel

$$k(\mathbf{X}_i, \mathbf{X}_j) = e^{-\frac{||\mathbf{X}_i - \mathbf{X}_j||^2}{2b}},$$

  and generate the kernel matrix $\mathbf{K}$ with the (i, j)th element being $k(\mathbf{X}_i, \mathbf{X}_j)$.
- We find a group of weights $\{w_i\}_{(i:D_i=0)}$ such that

$$\frac{1}{N_1} \sum_{i:D_i=1} k(\mathbf{X}_i, \mathbf{X}_j) \approx \sum_{i:D_i=0} w_i k(\mathbf{X}_i, \mathbf{X}_j)$$
$$\sum_{i:D_i=0} w_i = 1,$$

  for any j.
- $\{k(\mathbf{X}_i, \mathbf{X}_j)\}_{i \neq j}$ are seen as $N - 1$ covariates of i.

Kernel balancing

- Hazlett (2018) show that if

$$Y_i = \alpha + \tau_i D_i + \Phi(\mathbf{X}_i)'\beta + \varepsilon_i,$$
$$\langle \Phi(\mathbf{X}_i), \Phi(\mathbf{X}_j) \rangle = k(\mathbf{X}_i, \mathbf{X}_j),$$

  then these weights also satisfy

$$\frac{1}{N_1} \sum_{i:D_i=1} \Phi(\mathbf{X}_i) = \sum_{i:D_i=0} w_i \Phi(\mathbf{X}_i).$$

- For the Gaussian kernel, $\Phi(\mathbf{X}_i)$ encompasses all the continuous functions of $\mathbf{X}_i$ when $N \to \infty$.
- Similarly, the balancing stage introduces extra uncertainties which are often ignored in practice (Wong and Chan 2018).

Balancing: application

```
## Converged within tolerance

## The ebal ATT estimate is 2424.661

## The SE of ebal ATT estimate is 894.7984
```

References I

Aronow, Peter M, and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60 (1): 250–67.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis*, 25–46.

Hazlett, Chad. 2018. "Kernel Balancing: A Flexible Non-Parametric Weighting Procedure for Estimating Causal Effects." *Available at SSRN 2746753*.

Heckman, James J, Hidehiko Ichimura, and Petra E Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64 (4): 605–54.

Kline, Patrick. 2011. "Oaxaca-Blinder as a Reweighting Estimator." *American Economic Review* 101 (3): 532–37.

References II

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–65.

Ratkovic, Marc. 2019. "Rehabilitating the Regression: Honest and Valid Causal Inference Through Machine Learning."

Wong, Raymond KW, and Kwun Chuen Gary Chan. 2018. "Kernel-Based Covariate Functional Balancing for Observational Studies." *Biometrika* 105 (1): 199–213.

Xu, Yiqing, and Eddie Yang. 2021. "Hierarchically Regularized Entropy Balancing." *Available at SSRN 3807620*.

Zhao, Qingyuan, and Daniel Percival. 2016. "Entropy Balancing Is Doubly Robust." *Journal of Causal Inference*. De Gruyter.